

## Instructions

This assignment will require you to go through the exercises below using Stata (or any other statistical software you are familiar with). You can use Word to type the answers to the questions, and Word or Excel to produce the tables. Report each regression output for questions **2**, **2.(ii)**, **2.(iii)**, **3.(iii)**, **4** in a column of one single table (**Table 1**). Report coefficient estimates and standard errors in brackets, each with no more than 2 digits after the decimal point (unless the coefficient is very small, then you can rescale the variable or report more). Report 3 stars for statistical significance at the 1% level, 2 stars for 5%, and 1 star for 10%. Report Adjusted  $R^2$  and number of observations. Report your table following the format of this template table (note that the dependent variable can be different across columns), paying attention to how fixed effects are included:

**Table 1**

<b>Variables</b>	<b>Amount</b>	<b>Amount</b>	<b>Approval</b>
$X_1$	1.00*** (0.01)	1.00*** (0.01)	1.00*** (0.01)
$X_2$		1.00*** (0.01)	1.00*** (0.01)
$X_3$			1.00*** (0.01)
Area Fixed Effects	No	No	Yes
Adjusted $R^2$	0.10	0.10	0.10
N of Observations	10,000	10,000	10,000

No table or figure should extend across 2 pages. Every table and figure should have a title and a caption. The assignment should consist of 3 separate files. The first file is the **main document** including answers, tables, and graphs. This should have a maximum length of four A4 pages and a font size for the text of 12. The second file is the **Stata log file**, which can be produced using the `log` command in Stata, and should include the commands that you run and the Stata output. The third file is the **Stata do file**, which contains all the commands in Stata that you wrote to address all questions in the assignment. Your log file and do file should be accessible to an outside reader, incomprehensible or incomplete log files and do files will lead to point deduction.

## Questions

**[5 Points] for correct variables, graphs, tables, with appropriate labelling.**

You have recently started working at your country's Central Bank and your boss asked you to use your econometric skills to investigate the business model of FinTech lenders. You decide to focus on Bondora, one of the largest European FinTech lenders, as they provide extensive and free access to their data. You can find on Canvas the file "LoanData.zip", which contains the file "LoanData.csv" that can be directly imported into Stata. Note that a detailed description of all variables present in the dataset is available at this link <https://www.bondora.com/en/public-reports> you just need to scroll down to "Platform shared dataset legends" and open the scroll down menu called "Shared Legend Between Public and Private". Import the dataset in Stata using the command `insheet using LoanData.csv`.<sup>1</sup>

---

<sup>1</sup>Note that this dataset is different from the one of last year's assignment and will deliver different results.

1. [20 Points] Construct a table with descriptive statistics (command `tabstat`) for all the following variables: Applied loan amount (variable `appliedamount` in the data, measured in €), Granted loan amount (variable `amount` in the data, measured in €), Interest rate (variable `interest` in the data, measured in percentage points), Loan maturity (variable `loanduration` in the data, measured in months), Total applicant's income (variable `incometotal` in the data, measured in €), Total applicant's liabilities (variable `liabilitiestotal` in the data, measured in €), Married dummy (variable `maritalstatus` in the data, construct a dummy equal to 1 if the applicant is married or cohabitant, zero otherwise<sup>2</sup>), High risk dummy (variable `rating` in the data, construct a dummy equal to 1 if applicant's rating is "E", "F" or "HR", zero otherwise), New customer dummy (variable `newcreditcustomer` in the data, construct a dummy equal to 1 if applicant has no prior credit history on Bondora, zero otherwise). Make sure you rescale all variables measured in € to be measured in thousands of €.

In the table with descriptive statistics provide number of observations, mean, standard deviation, minimum, 10<sup>th</sup> percentile, median, 90<sup>th</sup> percentile, maximum. Include this table in your answer and briefly interpret these descriptive statistics. Plot two graphs on the relationship between interest rate and granted amount (command `twoway`), one for high risk and one for low risk applicants. Each of these graphs should only include two lines: a linear and a quadratic regression line for the relationship between the two variables. Make sure these two graphs have the same scale and labelling of the vertical axis, so that they are comparable, using the options `yscale` and `ylabel` within the `twoway` graph environment. For each of these two graphs explain whether you think the relationship between interest rate and amount is linear or non linear, and provide an economic interpretation for the shape of each curve.

2. [30 Points] Estimate an OLS regression model with interest rate as dependent variable, and granted loan amount, loan maturity, total applicant's income, total applicant's liabilities, married dummy, high risk dummy, and new customer dummy as independent variables. Report your results in **Table 1**.
  - (i) Interpret the statistical and economic significance of the estimated coefficients of the regression model you just estimated. Explain whether your findings make economic sense and why. Give a standardized coefficient interpretation to the effect of granted loan amount on interest rate. Interpret the goodness of fit of the model.
  - (ii) Based on the graphical evidence reported before, you decide to estimate the same regression model as in 2., but now with the only difference that you allow the effect of granted loan amount on interest rate to be different between high risk and low risk borrowers, using a multiplicative dummy. Report your results in **Table 1**. Interpret the economic and statistical significance of the effect of granted loan amount on interest rate for this new model, and argue whether and why your findings make economic sense.
  - (iii) Starting from the last model you estimated, include now in the regression fixed effects for borrowers' country of residence and for month when the loan was issued (again you will need to find these variables following the variables' description)<sup>3</sup>. Report your results in **Table 1**, without reporting the coefficient estimates for each of the new dummies, but following the example of the template table. Describe any relevant change in the

---

<sup>2</sup>Assign a zero also to the case of `maritalstatus=-1`.

<sup>3</sup>Months fixed effects require you to include a dummy for each month-year combination, therefore (for example) NOT a single dummy for each month of January across all years, but different dummies for January 2014, January 2015, etc...

model interpretation compared to your results in the previous question, both in terms of statistical and economic significance. Explain why it can be useful to include country and month fixed effects, and what the estimated coefficients for those might be capturing. Test the joint significance of the coefficients of the country fixed effects and, with a separate test, the joint significance of the month fixed effects. Report and interpret the result of the tests and decide whether those fixed effects jointly have a statistically significant effect on interest rate or not.

**3. [35 Points]** You now want to understand whether machine learning algorithms that detect potentially fraudulent borrowers provide a more effective screening of borrowers relative to manual checking. You decide to exploit the introduction of a new fraud detection technology, active on Bondora since October 2014, as potential treatment effect that improves the screening ability of the platform through machine learning algorithms. Consider then all months prior to October 2014 as months pre-treatment, and the months from October 2014 (included) onwards as months post-treatment. Consider as treated group the high risk borrowers, and as control group the rest. Use the loan interest rate as your dependent variable and a difference in differences estimator to quantify the effect of machine learning algorithms on the effectiveness of borrowers' screening.

- (i) Starting from the model you estimated in question 2.(iii), what assumption(s) and what controls do you need in this regression model to make sure that your difference in differences estimator is unbiased and consistent, and therefore captures the true causal effect of machine learning algorithms on the effectiveness of borrowers' screening?
- (ii) One important test to run before implementing a difference in differences estimation is checking whether the parallel trends assumption holds. Explain what this test is, why it is important, and what is the ideal outcome of it. Construct a graph in Stata using the command `graph twoway line`, where on the vertical axis there is the average monthly loan interest rate, and on the horizontal axis there are the months in the data, starting from January 2012. There should be two lines in this graph, one measuring the average monthly loan interest rate for high risk borrowers, and one measuring the average monthly loan interest rate for low risk borrowers. Moreover, you should include a straight vertical line corresponding to the month of October 2014. Produce this graph only for the borrowers of Estonia, and report the graph with all the correct labels (axis and legend). Based on this graph, argue whether and why in your context the parallel trends assumption holds or not.
- (iii) Starting from the model you estimated in question 2.(iii), implement a difference in differences estimation, report your results in **Table 1**, and give an economic and statistical interpretation only for the new elements of the model. What do you conclude regarding the effect of machine learning on borrowers' screening?

**4. [10 Points]** Last, you are interested in understanding the determinants of loan defaults. Keep only the loans that are either repaid or late (using the variable `status`), and estimate a probit model where the dependent variable is a dummy equal to one if the loan has defaulted and zero otherwise. You can construct this variable based on the variable `defaultdate`, assuming that observing no default date means that the loan hasn't defaulted. Use as independent variables the same ones used in question 2.(iii), but including only year fixed effects instead of month fixed effects. Report and interpret the average marginal effects, both in terms of

statistical and economic significance, for all variables except the fixed effects (where needed). Report your results in **Table 1**.